NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

TESTING FOR NORMALITY WITH THE KOLMOGOROV-SMIRNOV TEST

INTERNAL NOTE MSC-ED-IN-68-75

MANNED SPACECRAFT CENTER
HOUSTON, TEXAS

September 1968

INTERNAL NOTE MSC — ED-IN-68-75

TESTING FOR NORMALITY WITH THE KOLMOGOROV-SMIRNOV TEST

PREPARED BY

A. H. Feiveson

A. H. Feiveson

APPROVED BY

Jay M. Lewallen

Jay M. Lewallen
Chief, Theory and Analysis Office

Eugene H. Brock

Eugene H. Brock
Chief, Computation and Analysis Division

National Aeronautics and Space Administration
Manned Spacecraft Center
Houston, Texas

September 3, 1968

ABSTRACT

This paper provides a guide for using the Kolmogorov-Smirnov Goodness of Fit Test when testing for normality; especially in cases where the mean and variance must be estimated from the sample.

INTRODUCTION

An effective method for testing the hypothesis that a set of data comes from a specified distribution is the Kolmogorov-Smirnov Goodness of Fit Test. It has been shown (refs. 1 and 2) that if $[x_i]$ $i = 1,2,\cdots n$ represents a sample of n independent observations hypothesized to be from a population with cumulative distribution function $F(x)$, one may test this hypothesis by computing the "maximum deviation" statistic D given by

$$D = \max_x |F_n(x) - F(x)|$$

where $F_n(x) = \dfrac{\text{no. of observations less than or equal to } x}{n}$.

Under the assumed hypothesis, the distribution of D is independent of the function F , and its percentage points may be found in standard textbooks and tables, but with the restriction (not usually mentioned) that the assumed distribution function F must be completely specified. If F contains parameters which have to be estimated from the sample, the distribution of D is different than that given in standard tables, and in fact depends on F .

Lillifors (ref. 3) discusses the case where F is the normal distribution function with the mean and variance estimated from the sample, and gives percentage points of D as estimated by Monte Carlo runs of size 1000.

The purpose of this paper is to: 1) publicize the results of reference 1 with improved accuracy, 2) give percentage points of D for the normal case where the mean is known and the variance is unknown, and 3) point out the existence of a computer program KOLSMR (Kolmogorov-Smirnov) which may be used to perform the K-S test at MSC.

SYMBOL TABLE

$x_i (i=1,2,\cdots n)$	—	raw data
$x_{(i)}, x_{(k)}$ *ith*,*kth*	—	ordered data
n	—	sample size
F	—	theoretical distribution function
D	—	maximum deviation
$F_n(x)$	—	estimate of F based on sample of size n
α	—	probability
$C_{n\alpha}$	—	$1-\alpha$ percentage point of the null distribution of D for a sample of size n
$F_n(x-0)$	—	limit of $F_n(t)$ as $t \to x$ through values less than x
$F_n(x+0)$	—	limit of $F_n(t)$ as $t \to x$ through values greater than x

μ	—	known mean
y_i, z_i	—	standardized variates
$\overline{x}$	—	sample mean
s, s^*	—	sample standard deviations
PHI	—	standard normal distribution function
CDF	—	(cumulative) distribution function

MEAN AND VARIANCE UNKNOWN

The following table gives estimates and 99 percent confidence limits for percentage points of D under the null hypothesis of normality when the mean and variance must be estimated from the data. The estimates of the critical values, $C_{n\alpha}$, for sample sizes $n = 3, 4, 5, 10, 15, 20, 25,$ and 30 and significance levels $\alpha = .10, .05,$ and $.01$ are obtained by ordering 10 000 random values of D,··say $D_{(1)}$, $D_{(2)} \cdots {}_{(10\ 000)}$, and choosing the 10 000(1-α)*th* of the $D_{(i)}$; thus, if $\alpha = .05$, the estimate of $C_{n.05}$ is $D_{(9500)}$. These results, based on a larger number of runs than those in reference 1, have fairly good accuracy. Ninety-five percent confidence limits for $C_{n\alpha}$ are shown alongside the estimate.

TABLE I — CRITICAL VALUES OF D WHEN TESTING FOR NORMALITY WITH MEAN AND VARIANCE UNKNOWN.

10 000 Trials	α = .10			α = .05			α = .01		
n	99% Low	$\hat{C}_{n\alpha}$	99% High	99% Low	$\hat{C}_{n\alpha}$	99% High	99% Low	$\hat{C}_{n\alpha}$	99% High
3	.3659	.3672	.3689	.3748	.3760	.3770	.3826	.3831	.3835
4	.3402	.3447	.3497	.3702	.3741	.3774	.4064	.4098	.4148
5	.3148	.3183	.3213	.3394	.3435	.3474	.3873	.3928	.4015
10	.2399	.2424	.2454	.2617	.2642	.2684	.2988	.3035	.3096
15	.2005	.2022	.2051	.2177	.2205	.2245	.2531	.2589	.2646
20	.1740	.1756	.1773	.1891	.1915	.1942	.2188	.2236	.2274
25	.1579	.1594	.1612	.1712	.1730	.1753	.1973	.2014	.2058
30	.1450	.1468	.1483	.1576	.1590	.1612	.1837	.1885	.1920
over 30		$\frac{.805}{\sqrt{n}}$			$\frac{.86}{\sqrt{n}}$			$\frac{1.031}{\sqrt{n}}$	

$\hat{C}_{n\alpha}$ = estimate of $C_{n\alpha}$ based on 10 000 Monte Carlo runs where $C_{n\alpha}$ is such that $\Pr\{D > C_{n\alpha}\} = \alpha$ for a sample of size n .

MEAN KNOWN AND VARIANCE UNKNOWN

When the mean is assumed known, and the variance is unknown, simulations show that except for very small sample sizes ($n \leq 3$), the distribution of D is essentially the same as for the standard case where the mean and variance are known.[1] The standard table follows so that one might see the difference between this case and the unknown mean case (Table I).

Note that the critical values in Table II are considerably higher than in Table I. Consequently, if one erroneously uses Table II when the mean and variance are estimated, there is only a small chance of rejecting the null hypothesis even though it is false. For example, if the mean and variance are estimated from the sample, the probability of D exceeding .2205 for a sample of size 15 is about .05 (Table I). Thus one would reject a null hypothesis of normality at the 5 percent level if his observed values of D were greater than .2205. Hence, if Table II were used one would not reject the null hypothesis unless the observed D were greater then .338, an event with probability much less than .05.

[1] For $n = 3$, the exact values of $C_{3\alpha}$ for $\alpha = .10, .05$, and $.01$ are .659, .726, and .798 when the variance is unknown. For $n = 4$, the 99 percent confidence limits are (.556,.569), (.611,.632) and (.718,.741). Note that the standard values shown in Table II all lie in these confidence intervals. The same is true for larger values of n .

TABLE II. - CRITICAL VALUES OF D WHEN TESTING FOR NORMALITY WITH MEAN KNOWN AND VARIANCE KNOWN (STANDARD TABLE).

n	$\alpha = .10$ $C_{n\alpha}$	$\alpha = .05$ $C_{n\alpha}$	$\alpha = .01$ $C_{n\alpha}$
3	.642	.708	.829
4	.564	.624	.734
5	.510	.563	.669
10	.368	.409	.486
15	.304	.338	.404
20	.264	.294	.352
25	.240	.264	.317
30	.220	.242	.290
35	.210	.230	.270
over 35	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

THE COMPUTER PROGRAM KOLSMR

The Theory & Analysis Office in the Computation & Analysis Division at MSC has a computer program KOLSMR, (ref. 4), which performs the Kolmogorov-Smirnov test on a set of data. The program prints the maximum deviation D , and its critical values for either the case when the distribution F is completely specified, or when F is normal with unknown mean and variance.

Since the critical values are not distribution free when estimating parameters, they may be invalid if F is not normal. However, if no parameters are to be estimated, then the given values hold for any F .

Calling Sequence

The program is called by the following sequence:
CALL KOLSMR (X,N,F,KW,KN,D)

where: X is a singly dimensioned array of observations
N is the sample size
F is the theoretical distribution function
D is the maximum deviation, i.e.,

$$D = \max_x |F_n(x) - F(x)|$$

KW is an indicator giving the following modes of output:

KW < 0	The maximum deviation D and a table of critical values is printed.
KW = 0	Nothing is printed.

KW = $k (k = 1,2,\cdots)$ Every k*th* ordered observation $x_{(k)}$ is printed with the left and right limits of the estimated CDF, $F(x-0)$ and $F(x+0)$, the true CDF, $F(x)$, F and the maximum of the two differences:

$$|F(x_{(k)}) - F_n(x_{(k)} - 0)|$$

and

$$|F(x_{(k)}) - F_n(x_{(k)} + 0)|$$

with the appropriate sign of that difference.

Finally, the maximum deviation D is printed with a table of critical values.

KN is an indicator which functions as follows:

1. When the theoretical distribution function F (normal or otherwise) is completely specified, KN should be set equal to 0.

2. If one desires to test for normality with a known mean μ, but unknown variance, form the new observations $y_i = x_i - \mu$, set KN = 1, and let F be the standard normal CDF.

The program will compute standardized variates

$$Z_i = \frac{y_i}{S^*} = \frac{x_i - \mu}{S^*}$$

where

$$S^* = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2}$$

and test them against F .

3. If one desires to test for normality with both mean and variance unknown, set KN = 2 and test against the standard normal CDF.

The program will compute the standardized variates

$$Z_i = \frac{x_i - \overline{x}}{S}$$

where

$$\overline{x} = \frac{1}{n}\sum x_i$$

and

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

The Z_i will then be tested against the standard normal CDF.

EXAMPLE

Figure 1 is an example of the output from KOLSMR. In this example, 40 random numbers, uniformly distributed between 0 and 10, were tested for normality without specifying the mean and variance. The numbers were stored in an array X , and KOLSMR was called as follows:
CALL KOLSMR (X,40,PHI,1,2,D).

The function PHI is the standard normal CDF in the UNIVAC 1108 library and was declared EXTERNAL. Since it was desired to print every value of X , KW was set equal to 1, and since this was a test for normality with unknown mean and variance, KN was set equal to 2.

The columns K and $X_{(k)}$ are self-explanatory. Note that the $X_{(k)}$ appear in increasing order. The columns labeled FH(x-0) and FH(x+0) are the left-and right-hand limits as $x \to X_{(k)}$ of $F_n(x)$. This is equivalent to

$$\text{FH}(x-0) = \frac{k-1}{n}$$

$$\text{FH}(x+0) = \frac{k}{n}$$

For example, when $k = 17$, $\text{FH}(x-0) = \frac{16}{40}$, $\text{FH}(x+0) = \frac{17}{40}$,

$F(x) = \text{PHI}(Z_k) = \text{PHI}\left(\frac{X_{(k)} - \overline{X}}{S}\right) = \text{PHI}\left(\frac{5.527909 - 6.014982}{2.594344}\right) =$
.425539 .

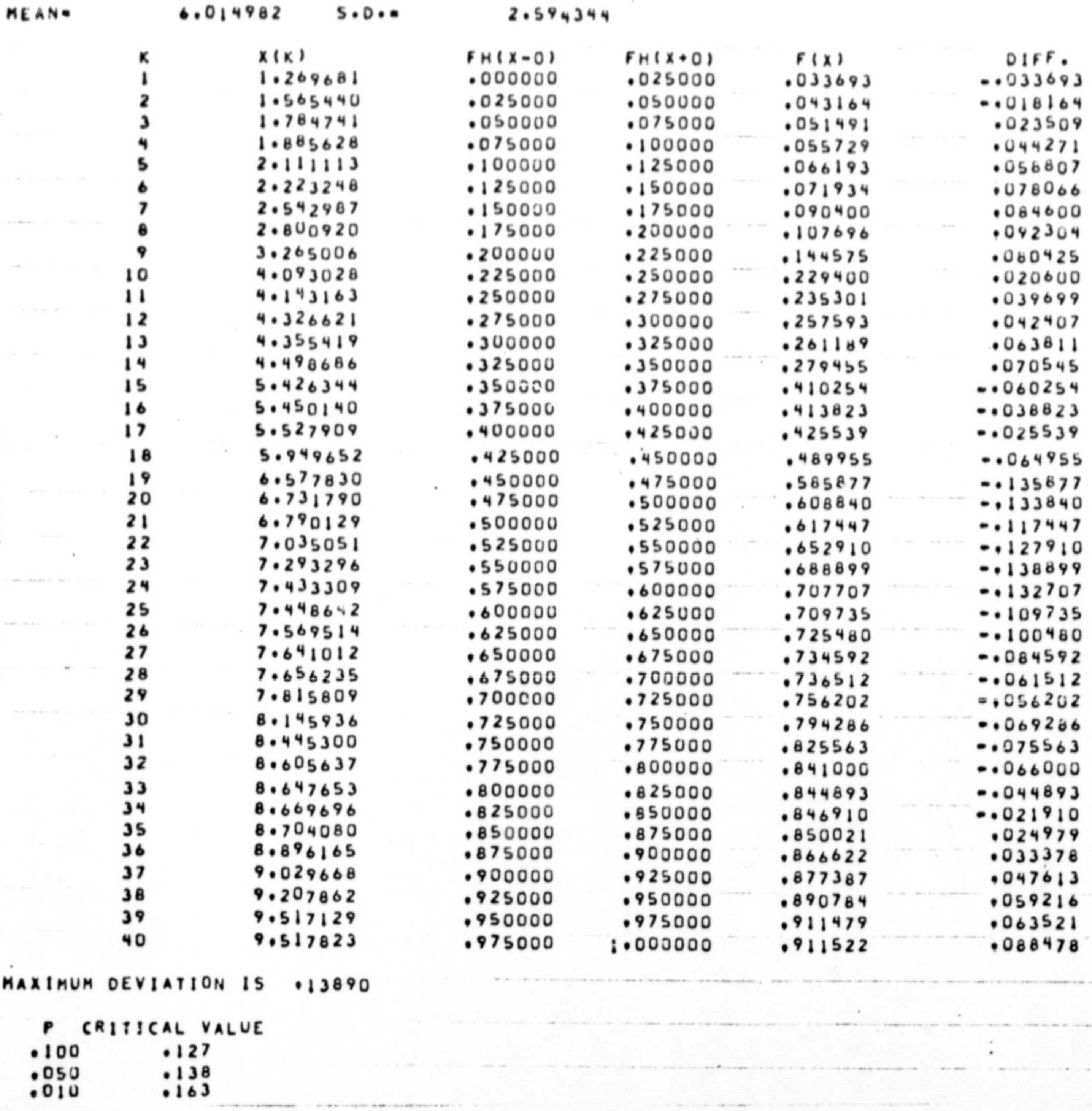
$\text{FH}(x+0) - F(x) = -.000539$, $\text{FH}(x-0) - F(x) = -.025539$.

DIFF is the above difference which is largest in magnitude; thus, DIFF = -.025539 .

The maximum deviation is the largest absolute difference; thus, D = |-.138899| (the maximum deviation occurs at observation number 23).

If the original data were normally distributed (the null hypothesis), the probability of D being larger than .138 is only .05. Since the observed value of D was .139, we therefore reject the null hypothesis of normality at the 5 percent level.

The sample mean and standard deviation ($\overline{x}$ and s) are printed at the top of the output if KN = 2 . If KN = 1 , the mean is always printed as 0. If KN = 0 , the mean is printed as 0, and the standard deviation is printed as 1.



```
MEAN=          6.014982     S.D.=           2.594344

            K          X(K)               FH(X-0)        FH(X+0)         F(X)                 DIFF.
            1          1.269681           .000000        .025000         .033693            -.033693
            2          1.565440           .025000        .050000         .043164            -.018164
            3          1.784741           .050000        .075000         .051491             .023509
            4          1.885628           .075000        .100000         .055729             .044271
            5          2.111113           .100000        .125000         .066193             .056807
            6          2.223248           .125000        .150000         .071934             .078066
            7          2.542987           .150000        .175000         .090400             .084600
            8          2.800920           .175000        .200000         .107696             .092304
            9          3.265006           .200000        .225000         .144575             .080425
           10          4.093028           .225000        .250000         .229400             .020600
           11          4.143163           .250000        .275000         .235301             .039699
           12          4.326621           .275000        .300000         .257593             .042407
           13          4.355419           .300000        .325000         .261189             .063811
           14          4.498686           .325000        .350000         .279455             .070545
           15          5.426344           .350000        .375000         .410254            -.060254
           16          5.450140           .375000        .400000         .413823            -.038823
           17          5.527909           .400000        .425000         .425539            -.025539
           18          5.949652           .425000        .450000         .489955            -.064955
           19          6.577830           .450000        .475000         .585877            -.135877
           20          6.731790           .475000        .500000         .608840            -.133840
           21          6.790129           .500000        .525000         .617447            -.117447
           22          7.035051           .525000        .550000         .652910            -.127910
           23          7.293296           .550000        .575000         .688899            -.138899
           24          7.433309           .575000        .600000         .707707            -.132707
           25          7.448642           .600000        .625000         .709735            -.109735
           26          7.569514           .625000        .650000         .725480            -.100480
           27          7.641012           .650000        .675000         .734592            -.084592
           28          7.656235           .675000        .700000         .736512            -.061512
           29          7.815809           .700000        .725000         .756202            =.056202
           30          8.145936           .725000        .750000         .794286            -.069286
           31          8.445300           .750000        .775000         .825563            -.075563
           32          8.605637           .775000        .800000         .841000            -.066000
           33          8.647653           .800000        .825000         .844893            -.044893
           34          8.669696           .825000        .850000         .846910            -.021910
           35          8.704080           .850000        .875000         .850021             .024979
           36          8.896165           .875000        .900000         .866622             .033378
           37          9.029668           .900000        .925000         .877387             .047613
           38          9.207862           .925000        .950000         .890784             .059216
           39          9.517129           .950000        .975000         .911479             .063521
           40          9.517823           .975000       1.000000         .911522             .088478

MAXIMUM DEVIATION IS  .13890

  P   CRITICAL VALUE
 .100       .127
 .050       .138
 .010       .163
```

Figure I. — EXAMPLE OF OUTPUT FROM KOLSMR

RESTRICTIONS

1. If one is testing data for a distribution other than normal, the theoretical CDF, F , must be completely specified.

2. The function F must be declared EXTERNAL in the calling program.

OTHER INFORMATION

The standard normal CDF PHI(X) is available to UNIVAC 1108 users on the system library at MSC. The Theory and Analysis Office also has decks of other CDF's, e.g., gamma beta, "t", which may be used on the 1108.

REFERENCES

1. Kolmogorov, A. N. "Sulla determinazione empirica di una legge distribuzione" – Giornale dell'Istituto Italiana Attuari, 4 (1933) pp. 83-91.

2. Darling, D. A. "The Kolmogorov-Smirnov, Cramér-von Mises Tests" – Annals of Mathematical Statistics 28 (1957) pp. 823-838.

3. Lillifors, Hubert W. "On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown" – J.A.S.A. June 1967, Vol. 62, p. 399.

4. Lockheed Program Library, Houston Aerospace Systems Division, Lockheed Electronics Co., Catalog No. 162, Program No. 5.1.